Protein Engineering vol.6 no.3 pp.327–331, 1993

Recursive ensemble mutagenesis

Exhibit 7

Simon Delagrave, Ellen R.Goldman and Douglas C.Youvan[1]

Massachusetts Institute of Technology, Department of Chemistry, Room 56-213, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[1]To whom correspondence should be addressed

We have developed a generally applicable experimental procedure to find functional proteins that are many mutational steps from wild type. Optimization algorithms, which are typically used to search for solutions to certain combinatorial problems, have been adapted to the problem of searching the 'sequence space' of proteins. Many of the steps normally performed by a digital computer are embodied in this new molecular genetics technique, termed recursive ensemble mutagenesis (REM). REM uses information gained from previous iterations of combinatorial cassette mutagenesis (CCM) to search sequence space more efficiently. We have used REM to simultaneously mutate six amino acid residues in a model protein. As compared to conventional CCM, one iteration of REM yielded a 30-fold increase in the frequency of 'positive' mutants. Since a multiplicative factor of similar magnitude is expected for the mutagenesis of additional sets of six residues, performing REM on 18 sites is expected to yield an exponential (30 000-fold) increase in the throughput of positive mutants as compared to random $[NN(G,C)]_{18}$ mutagenesis.

Key words: light harvesting II/protein engineering/random mutagenesis

Introduction

Current endeavors to engineer new specificities in antibodies and their derivatives hold the promise of new therapeutic and diagnostic tools. The generation of new and informative mutant proteins is necessary to our understanding of protein structure–function relationships. Such tasks are made difficult by our inability to predict structure from primary sequence or even to predict function from structure. One strategy circumventing the gaps in our understanding involves the selection of desired phenotypes from a large pool of different genotypes, in a manner analogous to natural selection. A limitation of this is the combinatorial explosion problem: as the number of randomized (mutated with all 20 amino acids) sites in a protein increases, the number of possible combinations which must be evaluated to identify 'positives' grows exponentially as 20^n, where n is the number of sites altered. Ingenious methods have been devised to allow screening of increasingly complex libraries of mutant proteins, peptides and oligonucleotides. Phage display libraries (Smith, 1985; Hoogenboom *et al.*, 1991; Kang *et al.*, 1991) and mutated ribozyme populations (Beaudry and Joyce, 1992) are instances of 'systems' where the genotypes and phenotypes are physically linked to allow for rapid selection and amplification of extremely complex ensembles of mutants. To completely screen a library of mutant proteins with 20 randomized amino acid residues ($n = 20$), the synthesis of 20^{20} or 10^{26} different protein molecules is required. Obviously, this will challenge our technical capabilities for some time. It may be desirable to avoid the very high proportion of non-functional proteins in a random library and simply enhance the frequency of functional proteins, thus decreasing the complexity required to achieve a useful sampling of sequence space. Recursive ensemble mutagenesis (REM) is an algorithm which enhances the frequency of functional mutants in a library when an appropriate selection or screening method is employed (Arkin and Youvan, 1992a; Youvan *et al.*, 1992).

REM uses successive rounds of CCM (Oliphant *et al.*, 1986; Reidhaar-Olson *et al.*, 1991) to generate a diverse library of genetically altered proteins that fit certain selection criteria (Figure 1). Amino acids are retained in the library if they are found in an altered protein fitting the selection criteria. Lists of all amino acids that are acceptable at each mutated position (i.e. 'target sets' of amino acids) are compiled. In the next iteration of REM, combinatorial cassettes are resynthesized according to mathematical functions that bias the nucleotide mixtures (Arkin and Youvan, 1992b; Youvan *et al.*, 1992) at each mutated

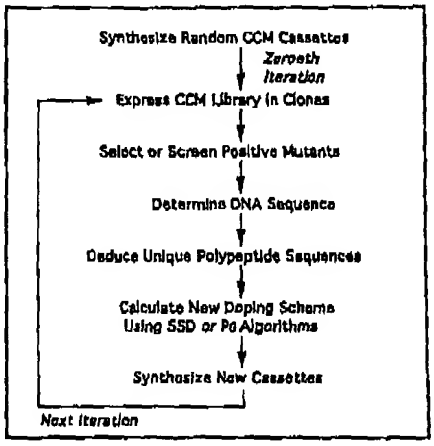

Fig. 1. REM involves the recursive use of combinatorial cassette mutagenesis (CCM). The first step of REM begins by expressing and screening a CCM library. Two or more 'positive' mutants are then picked and sequenced. (Positive mutants are defined in the current experiment as binding significant levels of red-shifted BchI which is characteristic of LHII assembly.) Next, a list of unique protein sequences is determined by translating these DNA sequences. A 'unique sequence' is defined at the protein level. If more than one protein has the same sequence, only the first occurrence of this sequence is retained and counted as unique. For each mutated position in the protein, a target set of acceptable amino acids is compiled and the most appropriate dope is determined by a mathematical function such as group probability (P_G). The next iteration of REM proceeds by using these 'intelligent' dopes to generate a combinatorial cassette of lower complexity. In order to take advantage of the properties of REM, the complexity of the possible peptide sequences arising from CCM should be shown to be in vast excess of the screening size (Youvan *et al.*, 1992).

position in the protein to encode these target sets of amino acids. For example, if Ala, Ser and Thr occur at a given position in different selected mutants, these amino acids constitute the target set at that position. A mathematical function is used to select the best 'dope' that maximizes the probabilities of the amino acids in the target set. The next cassette is then designed such that this target set is encoded by a simple mixture of nucleotides at that codon [e.g. [(G,A,T)(C)(G,C)]]. In certain cases, where there is a good match between selection criteria and structure inherent in the genetic code (Sjostrom and Wold, 1985; Youvan, 1991) such as hydropathy and molar volume, computer simulations predict that multiple iterations of REM will yield thousands of times more mutants than conventional CCM (Arkin and Youvan, 1992b; Youvan *et al.*, 1992).

As a model system to experimentally verify the computer predicted amplification by REM, the light harvesting II (LHII) β-subunit gene (Youvan and Ismail, 1985) of *Rhodobacter capsulatus* was chosen. The LHII protein has two characteristic absorption bands in the near infrared (800 and 858 nm) that are red shifted relative to protein-free bacteriochlorophyll (Bchl) absorption at 770 nm. These prosthetic groups serve as colorimetric indicators of protein expression and subunit assembly. Six carboxy-terminal residues of the β-subunit were initially mutated by construction of a combinatorial cassette containing the sequence $[NN(G,C)]_6$, where 'N' designates an equiprobable mixture of all four nucleotides. This CCM library was conjugated into a strain of *R.capsulatus* (U71) totally deficient in Bchl-binding proteins or any other compounds with significant absorption in the near infrared (Youvan *et al.*, 1985). This deletion background facilitates the use of digital imaging spectroscopy (DIS) (Arkin *et al.*, 1990; Arkin and Youvan, 1993) to screen thousands of colonies directly on Petri dishes for LHII expression. We then sequenced five functional mutants and used this limited data to construct a new CCM library. The frequency of positives was increased 30-fold relative to the original library.

Materials and methods

Plasmids and strains

Plasmid pU4b is a shuttle vector used for cassette mutagenesis as well as expression of the mutant LHII genes (Goldman and Youvan, 1992). M13 was our vector for single-stranded sequencing and was propagated in *Escherichia coli* MV1190. *Escherichia coli* strain S17-1 was used for library construction and conjugation with *R.capsulatus* U71. For expression of the libraries, *R.capsulatus* U71, an LHII chromosomal deletion background (LHII and reaction center expression inactivated by a point mutation) was used.

Materials and DNA manipulations

DNA manipulations were essentially performed as described by Sambrook *et al.* (1989). Restriction enzymes were obtained from New England Biolabs, T4 DNA ligase was from Bethesda Research Labs as was Taq polymerase. Sequencing was carried out using a Sequenase kit from United States Biochemicals. Electroporation was carried out in 0.2 cm cuvettes on 0.45 ml of competent cells using a Bio-Rad electroporator according to instructions provided. All oligonucleotides were synthesized on an Applied Biosystems model 381 DNA synthesizer using commercially available reagents.

Library construction

The unique *Kpn*I and *Xho*I sites of pU4b flank the region encoding the dimer Bchl binding site and the carboxy-terminus of the β-subunit LHII gene. These restriction sites were engineered to allow double-stranded combinatorial cassettes to be subcloned in place of the wild type sequence.

Table 1. Sequences and corresponding phenotypes of mutants isolated from the zero and first iterations of REM

Mutant	Deduced sequence	OD 800 nm	OD 858 nm	OD858/OD800
Wild type	ATPWLG	0.26	0.39	1.5
REM0.6	LTPWVA	0.15	0.24	1.6
REM0.7	LTPWVP	0.13	0.20	1.5
REM0.8	ASPWMS	0.09	0.15	1.7
REM0.9	SSPWLP	0.15	0.22	1.5
REM0.10	FVWPGF	0.05	0.09*	1.8
REM1.1	STPWVF	0.11	0.17	1.5
REM1.2	FTPWVG	0.11	0.18	1.6
REM1.3	ATPWLA	0.10	0.15	1.5
REM1.4	STPWLA	0.33	0.48	1.5
REM1.5	LTPWGR	0.09	0.13	1.4
REM1.6	VTPWLP	0.11	0.18	1.6
REM1.7	VTPWLG	0.13	0.21	1.8
REM1.8	LIWPVL	0.05	0.09*	1.8
REM1.9	ALWPLV	0.05	0.09*	1.8
REM1.10	LTPWGG	0.20	0.29	1.5
REM1.11	VTPWVR	0.06	0.11	1.8
REM1.12	VTPWGL	0.12	0.21	1.8

*Peak shifted to 845 nm.

The sense strand of the 113-mers, which included the *Kpn*I–*Xho*I sites, as well as two PCR primers (20-mers each spanning a restriction site) were synthesized. The doped sequence within the cassette used in the zero iteration was $[NN(G,C)]_6$. The purified 113-mer was amplified by PCR. Amplified double-stranded cassette was then purified by phenol extraction and ethanol precipitation. Complete digestion of the cassette with *Kpn*I and *Xho*I is carried out in a single incubation. The digested cassette is then purified by phenol and ether extractions and ultrafiltration in a Centricon 30 device (Amicon).

Ligation is carried out for 24 h at 16°C in 20 μl with approximately 0.1 μg of pU4b similarly digested with *Kpn*I and *Xho*I. The resulting pU4b derivative (an aliquot of the ligation) are directly electroporated into S17-1 *E.coli*. Aliquots of the transformation are plated on LB-tetracycline plates (after allowing 1 h for resistance expression) for complexity estimation and the remainder of the transformation is incubated overnight in 60 ml of LB-tetracycline. Plasmid pU4b derivatives were conjugated from *E.coli* S17-1 donors into *R.capsulatus* strain U71. The library is expressed by U71 transconjugants selected for by growth on RCV-tetracycline plates at 32°C.

Dope optimization

In computer simulations, various functions were used to optimize the 'nucleotide mixtures'. In this work, only five functional mutant sequences were obtained in the zero iteration. Given this small number of sequences and in order to conserve diversity, we elected to use the group probability (P_G) function because it retains all amino acids in the target set. When presented with a target set at one position, the program 'CyberDope' (provided courtesy of KAIROS Inc., Cambridge, MA, USA) goes through all integer nucleotide mixtures possible for a codon and evaluates for each mixture the value of P_G:

$$P_G = \prod_i P_D[i] \qquad (1)$$

where $P_D[i]$ is the frequency of occurrence of the ith amino acid

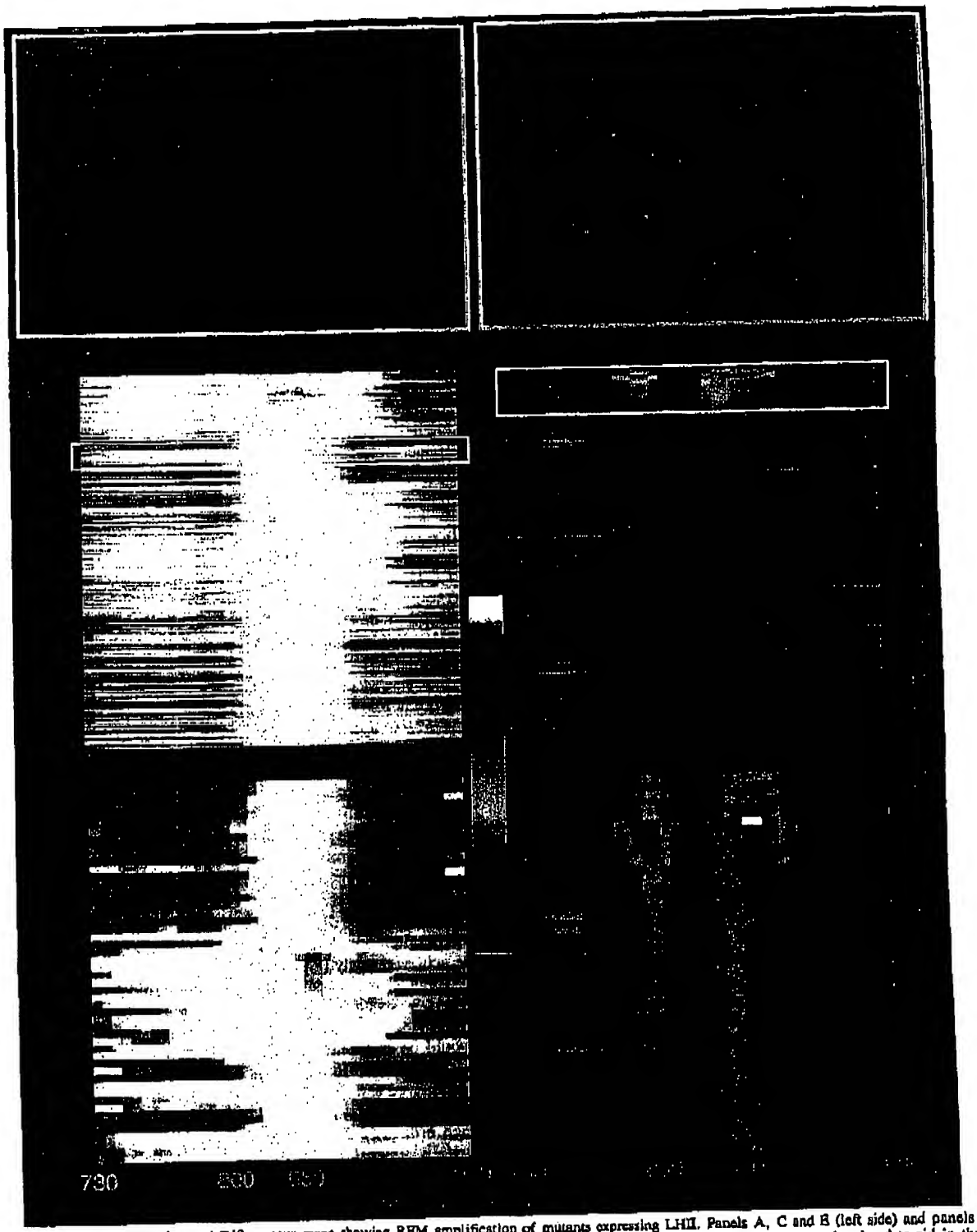

Fig. 2. Fluorescence screening and DIS contour maps showing REM amplification of mutants expressing LHII. Panels A, C and E (left side) and panels B, D and F (right side) correspond to the zero and first iterations of REM respectively. Fluorescence assays (panels A and B) were pseudocolored to aid in the visualization of relative levels of fluorescence according to the color bar shown below. These show brighter and more numerous fluorescent colonies in the first iteration of REM (panel B) than in the zero iteration of REM (panel A). DIS color contour maps (panels C and D) of the ground state absorption spectra of the same Petri dishes shown in panels A and B confirm the presence of mutants expressing Bchl-binding proteins. Each horizontal line represents the color coded absorption spectrum of a single colony; the color bar spans OD values of 0.0876–0.1177 for panels C and E and 0.0927–0.2088 for panels D and F. The spectra were sorted (Arkin and Youvan, in press) according to their absorption at 850 nm. Eight hundred and ninety-seven spectra from the zero iteration plate and 451 spectra from the first iteration plate are displayed in panels C and D respectively. Panels E and F highlight the spectra of 50 colonies from panels C and D (within the white boxes) respectively. Panel F shows many more colonies with spectra characteristic of LHII than found in panel E.

(in a target set of i amino acids) as encoded by a specific triplet dope. For the hypothetical target set mentioned above (Ala, Ser and Thr), any mixture not encoding a member of the target set (e.g. $P_D[Ala] = 0$) will cause P_G to be zero. The mixture with the highest value of P_G will be selected for the dope at that position. The doped sequence within the cassette used in the first iteration of REM was
[(G,T)(C,T)(C,G)][(A,G,T)(C,T)(C,G)][(C,T)(C,G)G][(C,T)(C,G)G][(A,G,T)(G,T)G][(C,T,G)(C,T,G)C]

Imaging spectroscopy

Colonies were imaged as spreads on RCV-tetracycline plates from the bacteria resuspended after conjugation. The most recent configuration of the digital imaging spectrophotometer has been described (Arkin and Youvan, in press). For the fluorescence images, the Petri dishes were illuminated with broad-band blue–green light and an 830 nm long pass filter was placed in front of the CCD lens to obtain radiometrically calibrated monochrome images which were linearly mapped to pseudocolors after establishing the low and high gray scale values for both images.

Results

The experimental complexity (i.e. number of independently generated clones) of the 'zero iteration' $[NN(G,C)]_6$ library was approximately 45 000. The theoretical complexity of such a library at the nucleotide level is calculated as 32^6 (1.1×10^9) because there are 32 possible [NN(G,C)] codons; the experimental complexity is only a small fraction of this number. Preliminary screening used fluorescence, (Yang and Youvan, 1988) which is indicative of LHII assembly, to rapidly identify mutants expressing LHII. Mutants are then more closely evaluated by ground state absorption measurements using DIS. We observed a low frequency of highly fluorescent colonies in the zero iteration of REM (ca. one positive mutant in 10 000 colonies screened). Relative to wild type absorption, DIS showed a decrease in the optical density at 800 and 858 nm for these few positives.

Because of their rarity, only five positives were obtained from the zero iteration of REM. Four of these five mutants fit the selection criterion of displaying significant absorbance at 858 nm and another, REM0.10, had an interesting phenotype. The five positives were repurified and sequenced (Table I). The composition of a first iteration cassette was calculated by the computer program 'CyberDope', which generates DNA dopes that maximize the overall probability of the target set. To add diversity to the target set, the wild type sequence was also included. Therefore, while not taking frequency of occurrence into account because of the small sample size, for the first doped position the target set is F, S, A, L. The output of CyberDope at the nucleotide level gave the codon [(G,T)(C,T)(C,G)], which encodes amino acids A, S, V (0.25 probability of occurrence for each) and F, L (0.12 probability of occurrence for each). Valine is unavoidably encoded by this dope because of the structure of the genetic code.

Figure 2 demonstrates the amplification properties of the REM methodology as assayed by digital imaging spectroscopy using both fluorescence emission and ground state absorption imagery. The first iteration of REM yields a 30-fold increase in the frequency of enhanced fluorescence mutants (Figure 2 A and B). As compared to zero iteration REM data, DIS analysis of the first iteration library shows both an increase in the percentage of positive mutants (i.e. throughput) and an increase in protein levels as determined by the intensity of the BChl absorption bands (Figure 2 and Table I).

Twelve positive mutants were sequenced from the first iteration REM library. All of these mutants express unique peptide sequences that differ from wild type. Two mutants (REM1.8 and REM1.9) show a 10 nm blue shift in the 858 nm band. These blue-shifted mutants have an inversion of the Pro–Trp motif found in all 29 sequences in the known phylogeny (Zuber, 1988) of β-subunits. This phenotype was first observed in the zero iteration library (mutant REM0.10), but now finds itself amplified in the first iteration. Note also that REM0.10 contains the same Pro–Trp motif inversion.

Discussion

To show that computer simulations were accurate in their prediction of an increased throughput of positives, an LHII gene was iteratively mutagenized at its six carboxy terminal residues. From the zero iteration (CCM) data, target sets of amino acids were defined. A computer encoded algorithm generated a doped oligonucleotide which best represented the target set at each mutagenized position. Expression of this new library (the first iteration of REM) revealed a substantial amplification in the throughput of pseudo wild type mutants. From the zero iteration library where roughly 10 000 colonies were screened to identify one positive, we can now conveniently identify a new positive by screening only about 300 colonies. This corresponds to a 30-fold increase in overall throughput, suggesting that mutating 18 sites of similar stringency would yield a 30^3 or 27 000-fold increase in throughput over random mutagenesis using $[NN(G,C)]_{18}$.

The altered proteins obtained by combinatorial mutagenesis are not necessarily trivial variations of the wild type sequence. An inversion of a completely conserved motif was observed in some mutants. Therefore, the sequence data indicate that REM does not recapitulate the known phylogeny. Mechanistically, the simultaneous (experimental) randomization of six sites in a protein may have no analogy in nature.

In this work, experimental evidence is given that REM allows an efficient search of sequence space by producing mutant libraries with increased frequencies of selected 'positives'. Due to the high stringency of the region chosen for mutagenesis, only a small sequence database was available for the construction of the first iteration dope. In systems where large complexities can be achieved easily (e.g. phage display libraries), more sites can be mutated at once and more positives isolated, giving a more complex sequence database. As a consequence, other dope optimizing equations (Youvan *et al.*, 1992) could be used which would be better suited to yield large increases in throughput. Alternatively, different short stretches of amino acids could be randomized and the zero iteration data from these libraries pooled to produce a first iteration dope mutagenizing many more sites than ordinarily possible with CCM.

It is important to make the connection between our algorithmically-based doping schemes and protein engineering projects where CCM is currently being used. REM decreases the fraction of null mutants in the population, therefore more sites can be simultaneously mutated. Model experiments on LHII can be used to optimize REM methodology, including the nucleotide doping equations. While DIS is limited to screening about 10^6 colonies, phage display libraries (Smith, 1985; Hoogenboom *et al.*, 1991; Kang *et al.*, 1991) can be used to select mutants from libraries with complexities exceeding 10^9. Based on our preliminary experiments, we expect greater phenotypic diversity

after one iteration of REM. This means that stronger 'binders' can be isolated, which is the fundamental goal of the phage display methodology. The use of CCM to introduce additional diversity in antibody libraries has already proven a useful approach (Barbas *et al.*, 1992) and may well be enhanced by the use of our mutagenesis scheme. REM is the first optimization technique that can be used to address this problem and explore sequence space in a mathematically rigorous fashion.

Acknowledgements

This work was supported by NIH GM42645, DOE 9102-025, DOE DE-FG02-90ER20019, and by a grant from the Human Frontiers Science Program. S.D. is partially supported by the funds FCAR. E.R.G. is the recipient of an NIH Biotechnology Training Grant Award.

References

Arkin,A.P. and Youvan,D.C. (1992a) *Proc. Natl Acad. Sci. USA*, 89, 7811–7815.
Arkin,A.P. and Youvan,D.C. (1992b) *Biotechnology*, 10, 297–300.
Arkin,A.P. and Youvan,D.C. (1993) In Deisenhofer,J. and Norris,J.R. (eds), *The Photosynthetic Reaction Center*, Academic Press, New York, in press.
Arkin,A.P., Goldman,E.R., Robles,S.J., Coleman,W.J., Goddard,C.A., Yang,M.M. and Youvan,D.C. (1990) *Biotechnology*, 8, 746–749.
Barbas,C.F., Bain,J.D., Hoekstra,D.M. and Lerner,R.A. (1992) *Proc. Natl Acad. Sci. USA*, 89, 4457–4461.
Beaudry,A.A. and Joyce,G.F. (1992) *Science*, 257, 635–641.
Goldman,E.R. and Youvan,D.C. (1992) *Biotechnology*, 10, 1557–1561.
Hoogenboom,H.R., Griffiths,A.D., Johnson,K.S., Chiswell,D.J., Hudson,P. and Winter,G. (1991) *Nucleic Acids Res.*, 19, 4133–4137.
Kang,A.S., Barbas,C.F., Janda,K.D., Benkovic,S.J. and Lerner,R.A. (1991) *Proc. Natl Acad. Sci. USA*, 88, 4363–4366.
Oliphant,A.R., Nussbaum,A.L. and Struhl,K. (1986) *Gene*, 44, 177–183.
Reidhaar-Olson,J.F., Bowie,J.U., Breyer,R.M., Hu,J.C., Knight,K.L., Lim,W.A., Mossing,M.C., Parsell,D.A., Shoemaker,K.R. and Sauer,R.T. (1991) *Methods Enzymol.*, 208, 564–587.
Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Sjostrom,M. and Wold,A. (1985) *J. Mol. Evol.*, 22, 272–277.
Smith,G.P. (1985) *Science*, 228, 1315–1317.
Yang,M.M. and Youvan,D.C. (1988) *Biotechnology*, 6, 939–942.
Youvan,D.C. (1991) *Trends Biochem. Sci.*, 16, 145–149.
Youvan,D.C. and Ismail,S. (1985) *Proc. Natl Acad. Sci. USA*, 82, 63–67.
Youvan,D.C., Ismail,S. and Bylina,E.J. (1985) *Gene*, 38, 19–30.
Youvan,D.C., Arkin,A.P. and Yang,M.M. (1992) In Muenner,R. and Munderick,B. (eds), *Parallel Problem Solving from Nature*, 2, Elsevier Publishing Co., Amsterdam, pp. 401–410.
Zuber,H. (1990) In Drews,G. and Dawes,E.A. (eds), *Molecular Biology of Membrane-bound Complexes in Phototrophic Bacteria*, Plenum Press, New York, pp. 161–180.

Received on November 6, 1992; revised on December 17, 1992; accepted on December 22, 1992